# Data Science Capstone Project 2:

## MECHANICAL FITTING FAILURE CLASSIFICATION PROBLEM

Content:

1. Notebook for the Project:
https://github.com/psanghal/Springboard-Data-Science/blob/master/Capstone Project 2/Project Notebook %26 Data/Capstone Project 2-Mechanical Failure.ipynb
(**Please copy and paste the full link directly in the browser to view the notebook**)

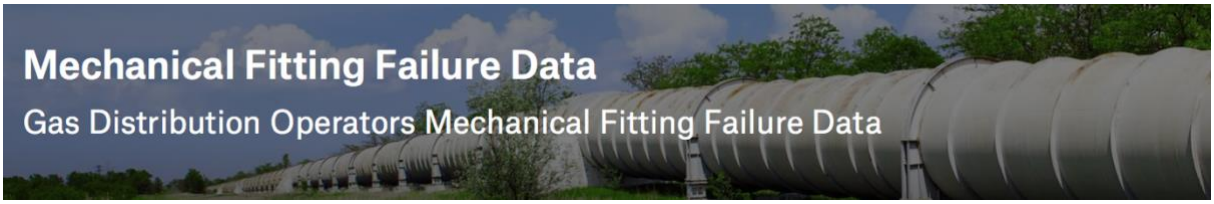
2. Presentation Slides Deck:
https://drive.google.com/file/d/1FJYKwK8z-ZFDlTe1wgpGVub06fpNjbJo/view?usp=sharing

3. Consolidated Report (See Enclosed)

*Prepared by:*
*Prashant Sanghal*

**Problem Statement:** Code of Federal Regulations (49 CFR Parts 191, 192) requires gas distribution pipeline operators to submit reports on an annual basis of all hazardous leaks that involve a mechanical fitting (DOT Form PHMSA F-7100.1-2).

Our goal is to classify mechanical fitting failure in the gas pipeline so that we can identify in- advance reasons that could have caused the leak to happen, when, and may be how often after the installation data?




If there are any fittings that failed during the first year of operation, then that issue could either be related to design or material defect caused by a third party, which should get reported to the CFR in a prompt manner.

**Benefit to Clients:** This report contains 326 pipeline operators and 225 manufacturers who are directly or indirectly involved in transporting gas to various locations using their pipeline network safely, reliably and long-term for their customers.

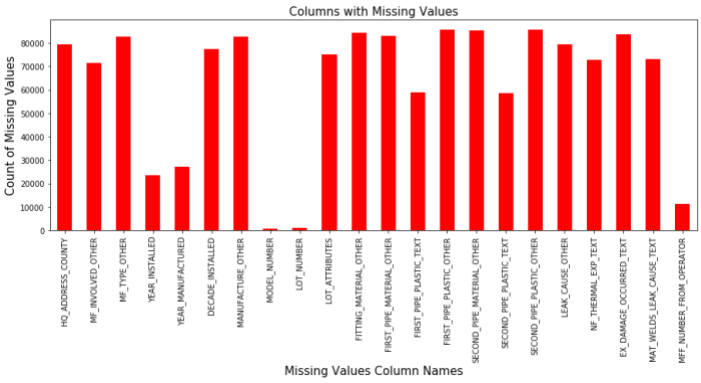
By training this model, we will help clients proactively manage their repair/ maintenance schedule on specific fittings that are impacted. Allow clients to engage manufacturers in research and development of long-life fittings. Set up early leak detection alert and avoid environmental hazards due to leak.

This report can also be used by insurance companies in estimating operator's insurance premium.

**Data Story:**

This dataset was available on Kaggle in comma seperated value, csv format. My mentor discovered this dataset and recommended me to work on it. This dataset contained 85,611 observations and 54 columns detailing various attributes about mechanical fittings used in the gas pipeline, some as old as 165 years old.

One of the major challenges with this dataset was that it was entirely in text format. As an example, failure date when the leak occurred, installation date when the fittings were installed and pipe nominal sizes, all were presented in the text format which had to be transformed into date time and numerical values.




Moreover, there were multiple columns where information regarding leak and the manufacturers were captured in two separate columns. One represented by column label ending with _TEXT while the other as a subset represented by _OTHER. The columns ending with _TEXT had multiple missing values which had to be replaced by values in the _OTHER columns.

To get a better understanding of the missing values, I wrote a function missing_dashboard which gave an overview on missing columns and unique values as shown in the table below.

The size of this table is dependent on a user defined variable 'missing_percentage' which can be selected anywhere between 0 to 1. For this table as an example, I chose 0.7 as the missing_percentage, which collected information on 15 missing columns out of 22 columns.

```
Average (base) Missing Percentage: 0.73
Number of Missing Columns
(Missing Percentage at 0.7): 15
```

| | missing_values | missing_percentage | unique_values_in_missing_columns | available_values_in_missing column |
|---|---|---|---|---|
| **HQ_ADDRESS_COUNTY** | 79345 | 0.93 | 84 | 6266 |
| **MF_INVOLVED_OTHER** | 71434 | 0.83 | 731 | 14177 |
| **MF_TYPE_OTHER** | 82557 | 0.96 | 317 | 3054 |
| **DECADE_INSTALLED** | 77247 | 0.90 | 10 | 8364 |
| **MANUFACTURE_OTHER** | 82683 | 0.97 | 196 | 2928 |
| **LOT_ATTRIBUTES** | 83343 | 0.97 | 723 | 2268 |
| **FITTING_MATERIAL_OTHER** | 84411 | 0.99 | 52 | 1200 |
| **FIRST_PIPE_MATERIAL_OTHER** | 82845 | 0.97 | 41 | 2766 |
| **FIRST_PIPE_PLASTIC_OTHER** | 85557 | 1.00 | 24 | 54 |
| **SECOND_PIPE_MATERIAL_OTHER** | 85274 | 1.00 | 53 | 337 |
| **SECOND_PIPE_PLASTIC_OTHER** | 85568 | 1.00 | 16 | 43 |
| **LEAK_CAUSE_OTHER** | 79267 | 0.93 | 778 | 6344 |
| **NF_THERMAL_EXP_TEXT** | 72766 | 0.85 | 2 | 12845 |
| **EX_DAMAGE_OCCURRED_TEXT** | 83503 | 0.98 | 2 | 2108 |
| **MAT_WELDS_LEAK_CAUSE_TEXT** | 73060 | 0.85 | 2 | 12551 |

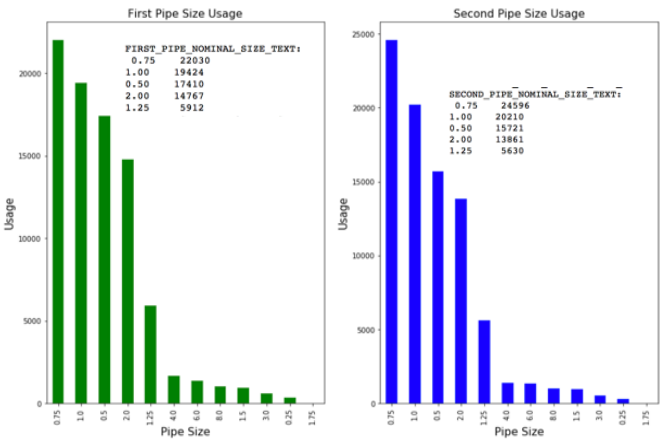
As we can see above, there are multiple unique values in the missing columns which can be used to replace missing values in other columns. To find this out, I wrote a function to help identify missing replacement values from _OTHER columns to fill missing values in the _TEXT column.

1. missing_value_replacement
2. missing_value_index_from_column_1
3. getvalues_from_column_2_using_missing_index_from_column_1
4. missing_replacement_value_and_index_column_2

As a result, we were able to replace a total of **18,789** missing values in a total of 9 columns.

**Exploratory Analysis:**

After replacing some missing values, my focus shifted on wrangling dates and getting insights from the data. I was interested in exploring which states in the US had maximum number of leaks by manufacturers. Which Manufacturers accounted for maximum number of leaks and what was the primary cause for the leak. Below, are some exploratory graphs which will help us answer these questions.

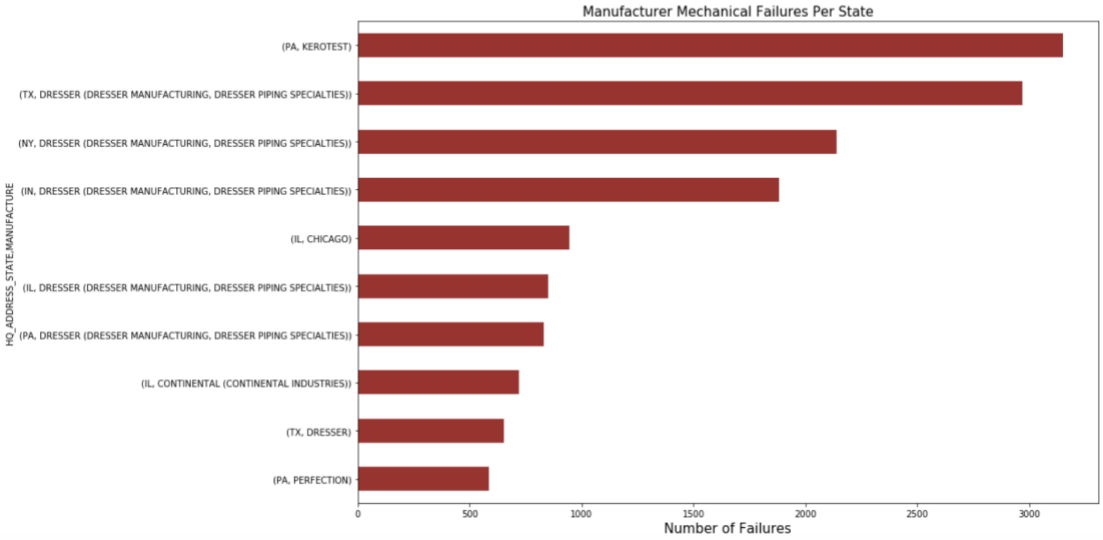



- **Most Used Pipe Sizes:**

After converting pipe size from text format to numerical values, pipe size of 0.75 showed most usage among pipeline operators with a transactional activity of 22k to 24K over the years.

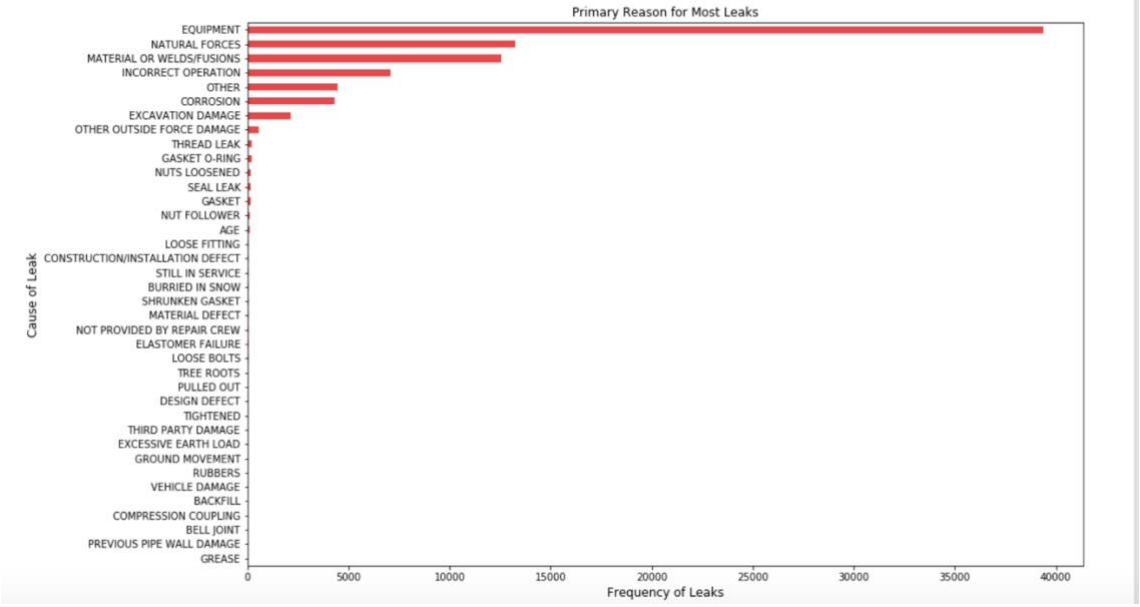
- **The Top 10 Known Manufacturers by State where Leak Occurred:**

```
HQ_ADDRESS_STATE  MANUFACTURE
PA                KEROTEST                                                  3152
TX                DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)  2972
NY                DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)  2138
IN                DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)  1881
IL                CHICAGO                                                    944
Name: MANUFACTURE, dtype: int64
```




- **Top 15 Reasons for the Leak:**

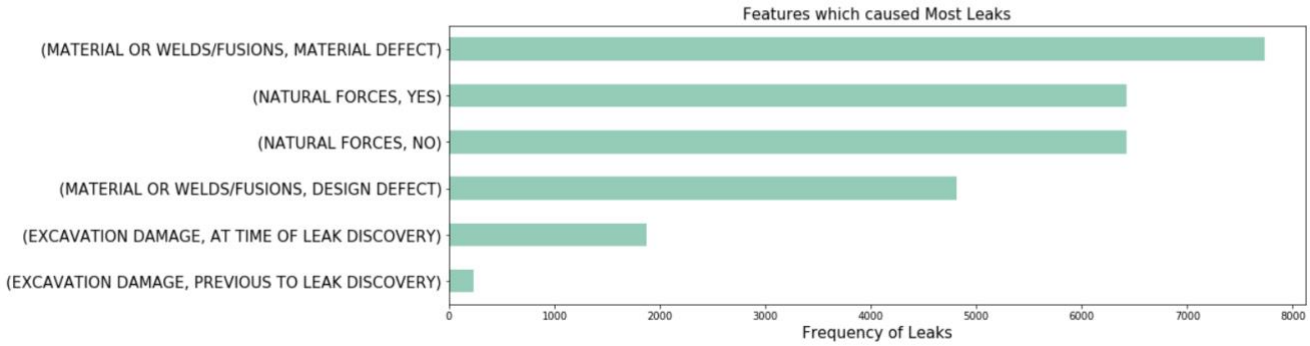
```
****Top 15 reasons for leak****
                            Reason_Count  % Reason_Count
EQUIPMENT                          39370       45.987081
NATURAL FORCES                     13230       15.453622
MATERIAL OR WELDS/FUSIONS          12551       14.660499
INCORRECT OPERATION                 7070        8.258285
OTHER                               4464        5.214283
CORROSION                           4330        5.057761
EXCAVATION DAMAGE                   2123        2.479822
OTHER OUTSIDE FORCE DAMAGE           575        0.671643
THREAD LEAK                          225        0.262817
GASKET O-RING                        223        0.260481
NUTS LOOSENED                        180        0.210253
SEAL LEAK                            163        0.190396
GASKET                               144        0.168203
NUT FOLLOWER                         133        0.155354
AGE                                   90        0.105127
```



- **Features which Caused Leak in the Pipeline:**

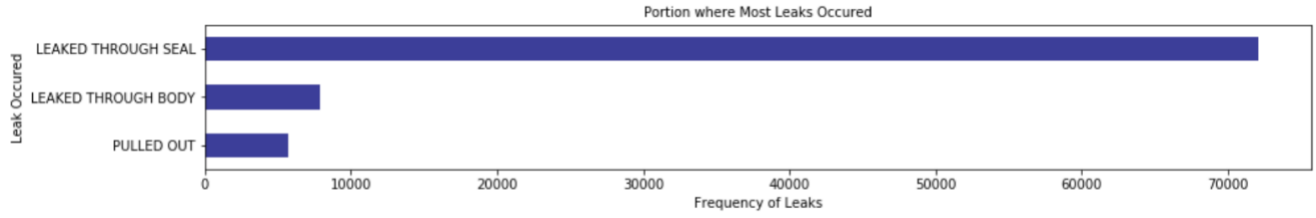
```
****Features which caused leak****
                                                    Feature_Count  \
LEAK_CAUSE_TEXT           ADDITIONAL_LEAK_FEATURES
MATERIAL OR WELDS/FUSIONS MATERIAL DEFECT                    7734
NATURAL FORCES            YES                                6424
                          NO                                 6421
MATERIAL OR WELDS/FUSIONS DESIGN DEFECT                      4817
EXCAVATION DAMAGE         AT TIME OF LEAK DISCOVERY          1872
                          PREVIOUS TO LEAK DISCOVERY          236

                                                    %_Feature_Count
LEAK_CAUSE_TEXT           ADDITIONAL_LEAK_FEATURES
MATERIAL OR WELDS/FUSIONS MATERIAL DEFECT                 9.033886
NATURAL FORCES            YES                             7.503709
                          NO                              7.500204
MATERIAL OR WELDS/FUSIONS DESIGN DEFECT                   5.626613
EXCAVATION DAMAGE         AT TIME OF LEAK DISCOVERY       2.186635
                          PREVIOUS TO LEAK DISCOVERY      0.275666
```




- **Portion where Most Leaks Occurred:**

```
****Portion where leak occured****
                    Occurred_Count  % Occurred_Count
LEAKED THROUGH SEAL          72062         84.173763
LEAKED THROUGH BODY           7867          9.189240
PULLED OUT                    5682          6.636998
```



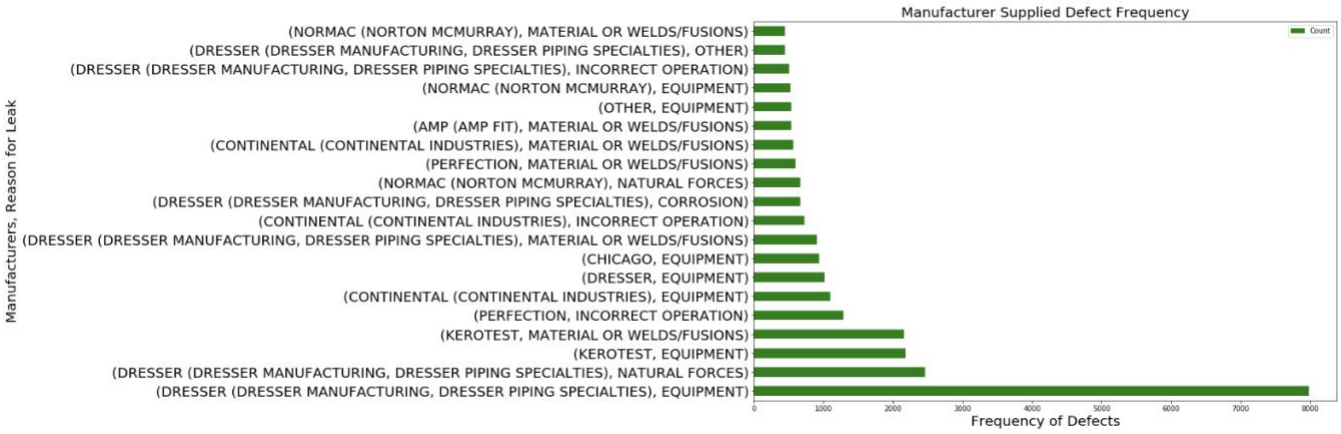

**Observation Summary:**

- 39.4K or 46% leaks were caused due to **equipment failure**, where features of failure were not known.
- 7.7k or 9% leak were caused due to **welding defects.**
- 72.1k or 84% **leaked through the seal**.

- **Manufacturer Supplied Defects:**

Manufacturer Supplied Defect Frequency:

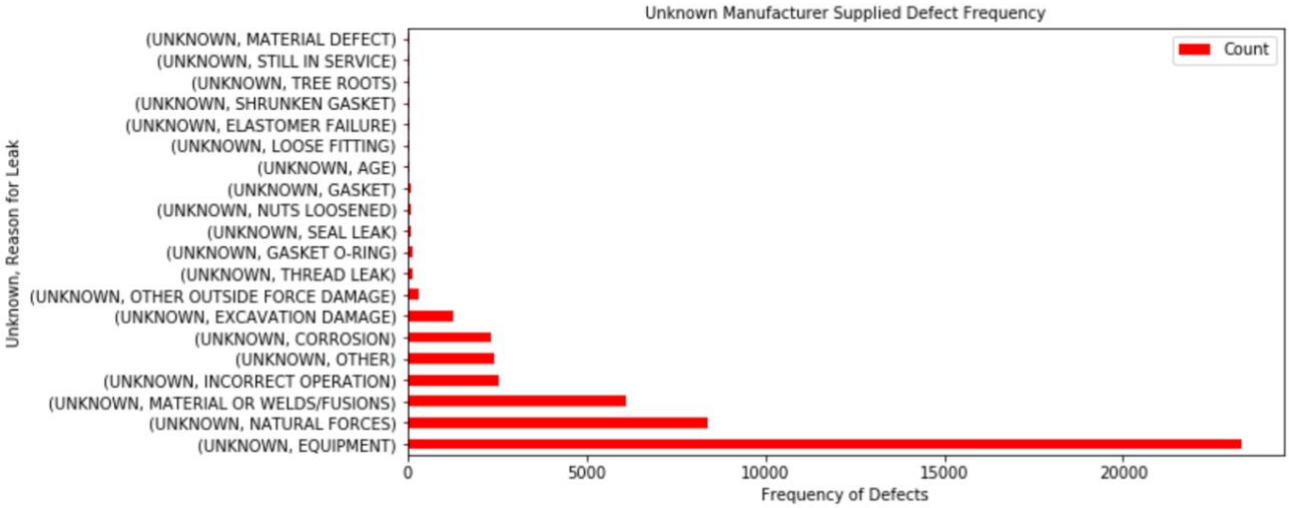
| MANUFACTURE | LEAK_CAUSE_TEXT | Count |
|---|---|---|
| DRESSER (DRESSER MANUFACTURING, DRESSER PIPING ... | EQUIPMENT | 7985 |
| | NATURAL FORCES | 2461 |
| KEROTEST | EQUIPMENT | 2177 |
| | MATERIAL OR WELDS/FUSIONS | 2162 |
| PERFECTION | INCORRECT OPERATION | 1285 |




- **Unknown Manufacturer Supplied Defects:**

Unknown Manufacturer Supplied Defect Frequency:

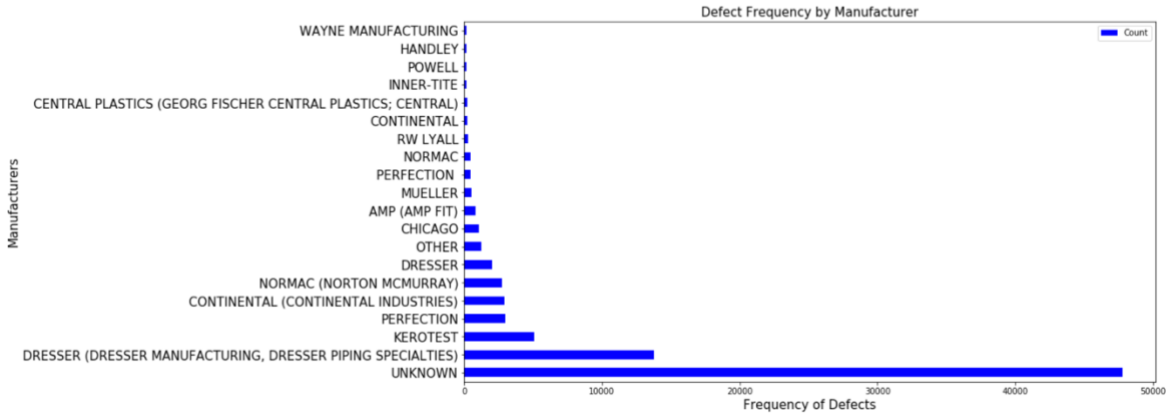
| MANUFACTURE | LEAK_CAUSE_TEXT | Count |
|---|---|---|
| UNKNOWN | EQUIPMENT | 23328 |
| | NATURAL FORCES | 8408 |
| | MATERIAL OR WELDS/FUSIONS | 6108 |
| | INCORRECT OPERATION | 2537 |
| | OTHER | 2434 |



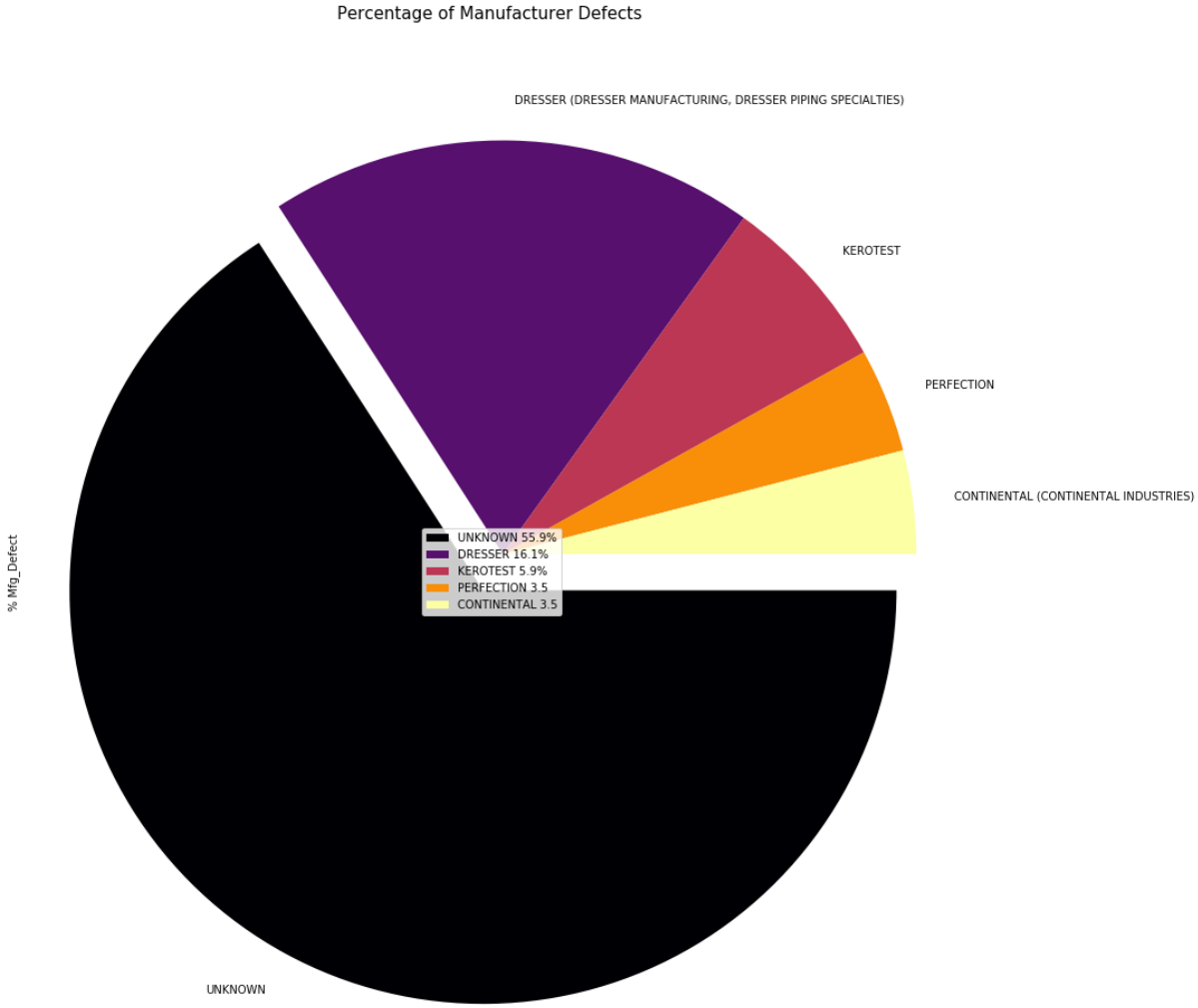
- **Defect Frequency by Manufacturers:**

Defect Frequency by Manufacturer:

| MANUFACTURE | Mfg_Count | % Mfg_Defect |
|---|---|---|
| UNKNOWN | 47816 | 55.887234 |
| DRESSER (DRESSER MANUFACTURING, DRESSER PIPING ... | 13808 | 16.138760 |
| KEROTEST | 5073 | 5.929311 |
| PERFECTION | 2954 | 3.452629 |
| CONTINENTAL (CONTINENTAL INDUSTRIES) | 2928 | 3.422240 |



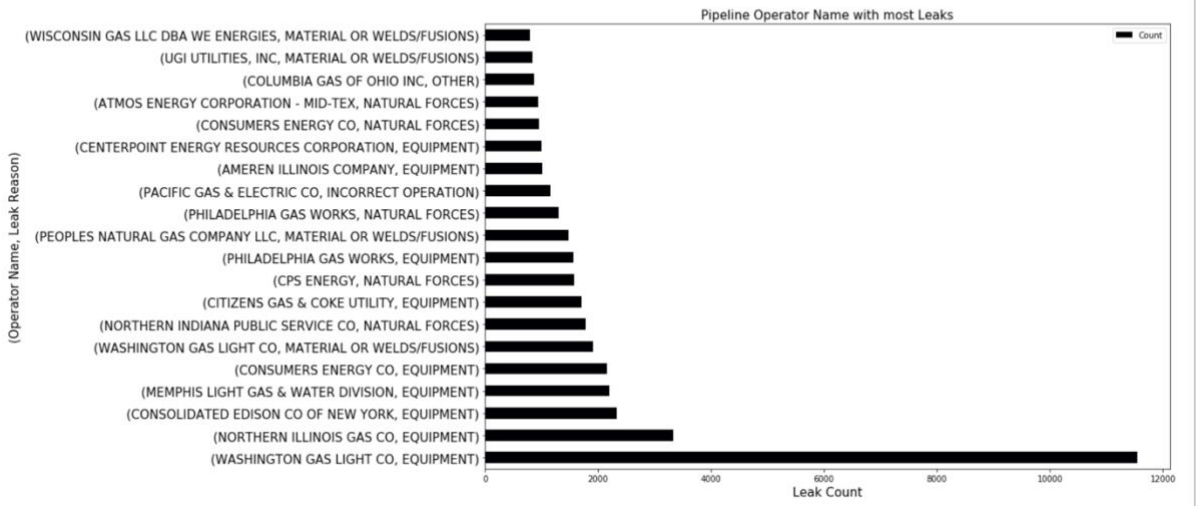

- **Percentage of Defects by Manufacturer:**



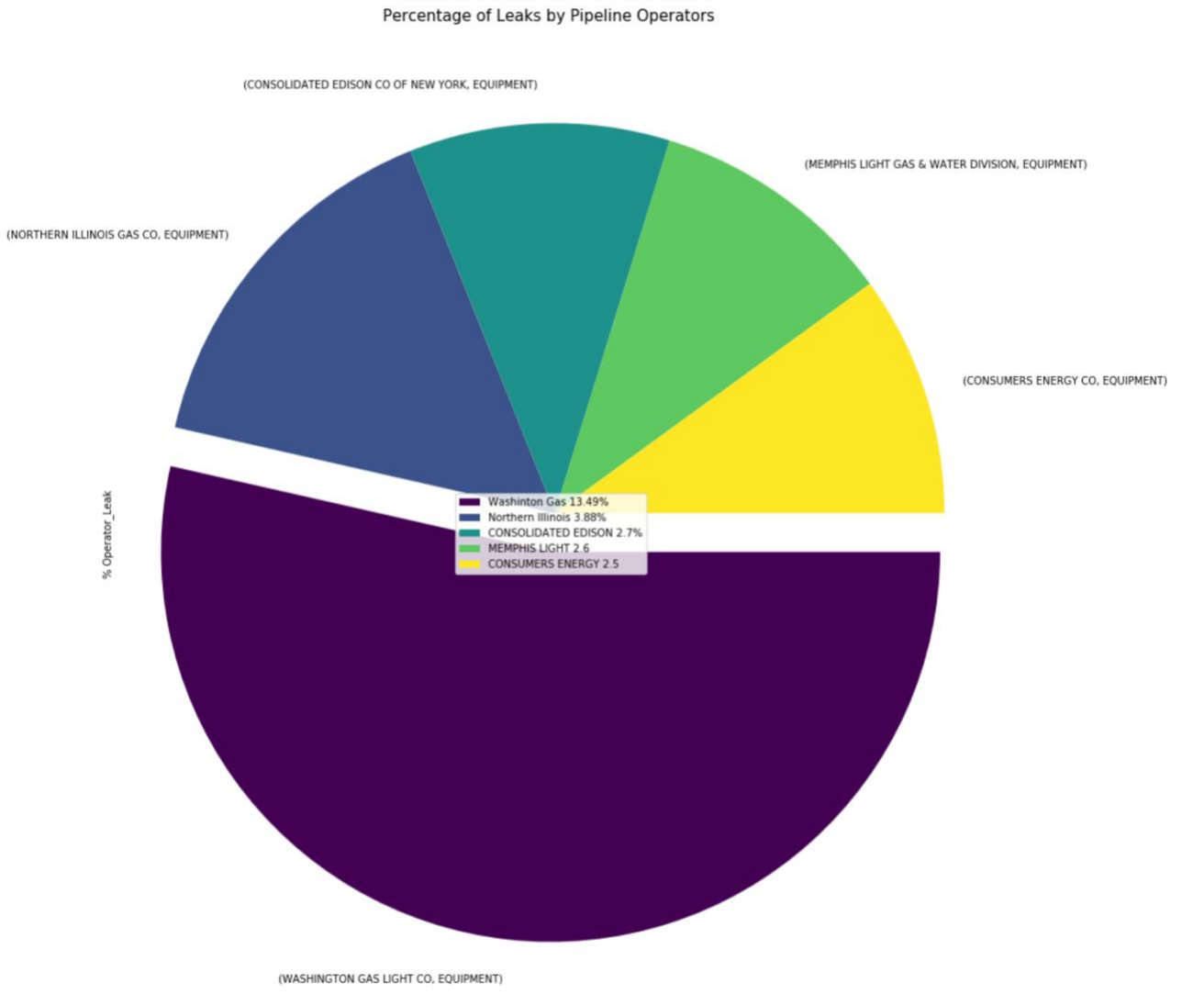
- **Pipeline Leaks by Operator:**

Leak Frequency by Pipeline Operator:

| OPERATOR_NAME | LEAK_CAUSE_TEXT | Operator_Count | % Operator_Leak |
|---|---|---|---|
| WASHINGTON GAS LIGHT CO | EQUIPMENT | 11551 | 13.492425 |
| NORTHERN ILLINOIS GAS CO | EQUIPMENT | 3326 | 3.885015 |




- **Percentage of leaks by Pipeline Operators:**



**Observation Summary :**

**Manufacturers:**

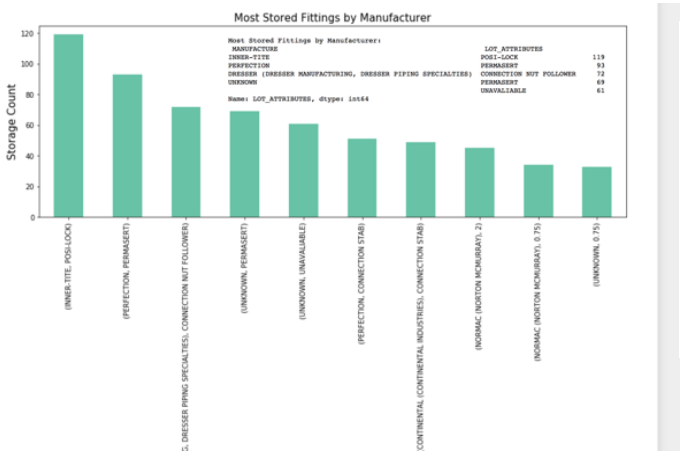
- 56% defects were caused by Unknown manufacturers, majority related to equipment failure.
- Dresser and Kerotest accounted for 16% and 6% defects related to equipment, natural forces and equipment, welding failures respectively.

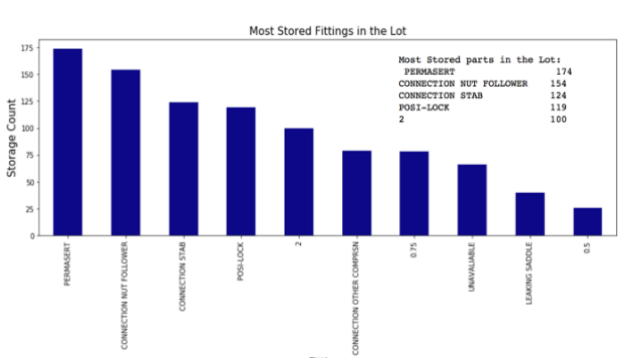
**Operators:**

- Most pipeline operator leaks were due to equipment failure.
- Washington Gas Light and Northern Illinois pipeline operators recorded 13.49% and 3.88% leaks respectively.

**Most Stored Fittings by Manufacturer and By Lot:**

After cleaning lot attributes and exploring manufacturers, Permasert was observed to be the most stored fitting in the lot supplied by two Dresser and Perfection.
While, per manufacturer basis Inter-Tite, the manufacturer of POSI-LOCK carried maximum fittings in the lot compared to other manufacturers.




**Relationship between dates (Installation Vs Report Vs Filing):**

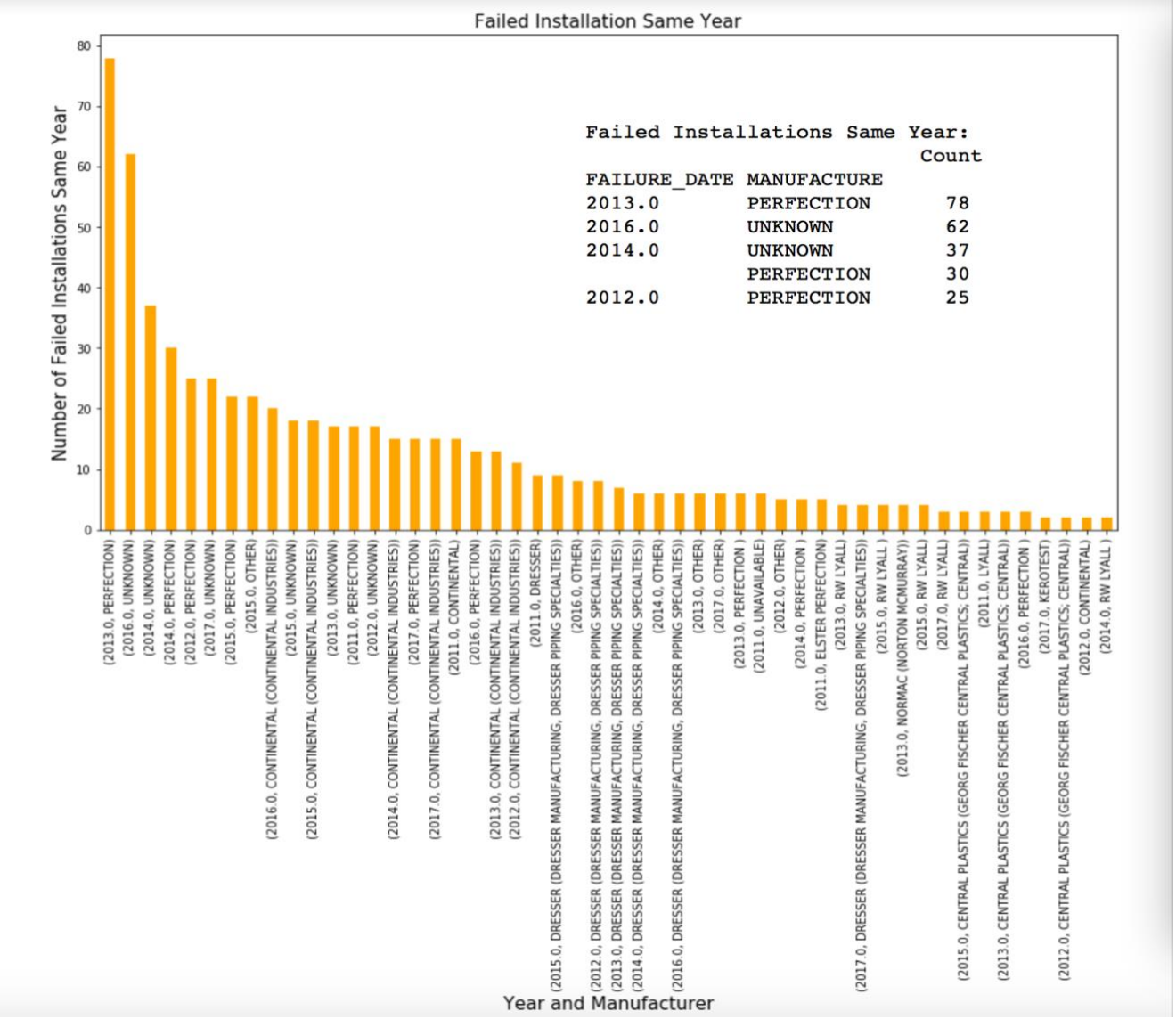
It was observed that  there was a significant delay  (12 years +) between failure occurred date wen cpmpared to reported and filed date.
After further investigation, it was clear that majority of the reports were filed the same day as the failure date. Hence, filing date was corrected and matched with the failure date to avoid any date lag in our study.

**Oldest Installation:**

First Mechanical fitting was installed 165 years ago in the year 1851. There were 6 installations done back then by Unknown manufacturer.s

**Same Year Failed Installation:**

Interestingly, there were also installations which failed within the first year of installation as shown in the graph below:




**Observation Summary:**

1. Perfection had maximum number of failed installations in 2013.
2. Dresser, Continental and Perfection had multiple failed installations between 2011 to 2017 as shown in the graph above.

After exploratory analysis, 20 columns were dropped and 4 new feature columns were added to extract meaningful insights from the raw data.

**Machine Learning:**

In the final phase after data pre-processing, to prepare our data for modeling, it was imporatnt to check for missing and categorical values and see if it requires any further treatment.

There were still multiple missing values present in the dataset which had to be topped up by frequently occurring values a.k.a mode values in the column or replaced with 'Unknown' where mode values presented a bias due to large amount of missing information.

A total of **155,047** values in various columns were replaced to completely avoid missing fields in the final table, df_cleaned.

In the last step before modeling, multiple categorical values had to be converted in to numerical values, which was done by using category encoding method discussed below:

1. One-hot-encoder.
2. Binary encoder.

The main difference between the two approaches was in terms of the data size it generated after conversion.

For example: One-Hot generated 3,912 columns after optimizing with label encoder while Binary Encoder generated 136 columns, which was significantly less when compared to one-hot ecoder.

In order to evaluate the difference between the two approaches, I modeled two data sets separately, Model A (one-hot) and Model B (Binary) before applying machine learning model.

**Model Selection:**

Given the data structure, I decided to evaluate the two models using Random Forest Classifier. Random forests can handle unbalanced data, outliers and non-linearity well. The only drawback however is that it can overfit the data which can be managed by hyperparameter tuning.

To identify best hyper parameters for the model, I started with Randomized Search CV and finished selection using Grid Search CV. During this process, we made 90 fits and it took 1 hour and 9 minutes to get the following hyper parameters using Model B data.

| Selected Hyper Parameters | Description |
|---|---|
| {'n_estimators': 100, | #Number of Trees |
| 'min_samples_split': 10, | #Minimum samples required to spilt internal node |
| 'min_samples_leaf': 1, | #Minimum samples at leaf node |
| 'max_features': 100, | #Maximum features to consider for split decision |
| 'max_depth': 50, | #Maximum depth of each tree |
| 'criterion': 'entropy'} | #Quality of split |

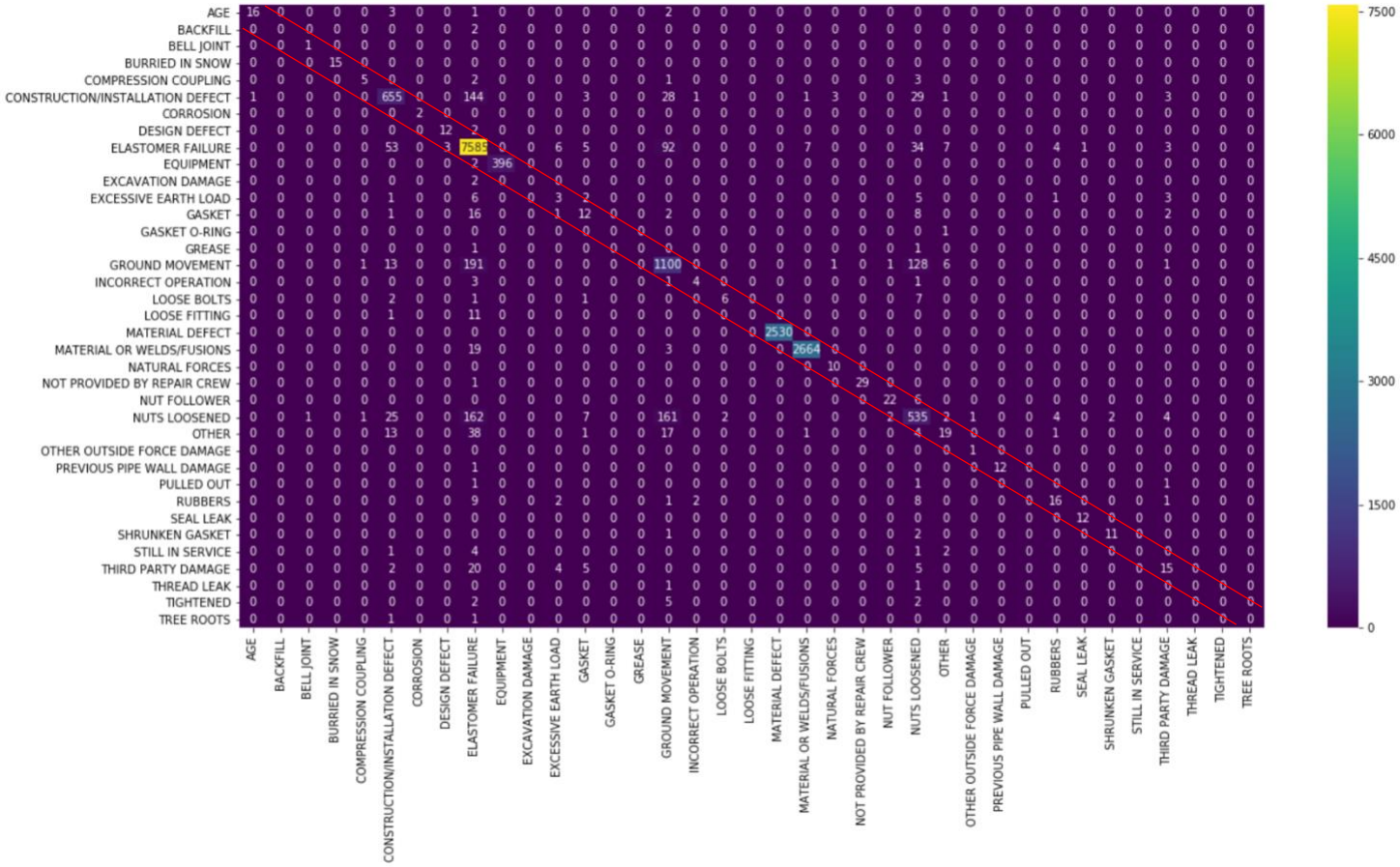
**Model Evaluation:**

**Model B (Binary Encoded):** As per classification report, we were able to train and classify 38 reasons that caused a mechanical fitting leak in the gas pipeline. Below is the classification report and a visual of confusion matrix, which will simply tell us how many correct and incorrect predictions for each leak label was made .

Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| AGE | 0.94 | 0.73 | 0.82 | 22 |
| BACKFILL | 0.00 | 0.00 | 0.00 | 2 |
| BELL JOINT | 0.50 | 1.00 | 0.67 | 1 |
| BURRIED IN SNOW | 1.00 | 1.00 | 1.00 | 15 |
| COMPRESSION COUPLING | 0.71 | 0.45 | 0.56 | 11 |
| CONSTRUCTION/INSTALLATION DEFECT | 0.85 | 0.75 | 0.80 | 869 |
| CORROSION | 1.00 | 1.00 | 1.00 | 2 |
| DESIGN DEFECT | 0.80 | 0.86 | 0.83 | 14 |
| ELASTOMER FAILURE | 0.92 | 0.97 | 0.95 | 7800 |
| EQUIPMENT | 1.00 | 0.99 | 1.00 | 398 |
| EXCAVATION DAMAGE | 0.00 | 0.00 | 0.00 | 2 |
| EXCESSIVE EARTH LOAD | 0.19 | 0.14 | 0.16 | 21 |
| GASKET | 0.33 | 0.29 | 0.31 | 42 |
| GASKET O-RING | 0.00 | 0.00 | 0.00 | 1 |
| GREASE | 0.00 | 0.00 | 0.00 | 2 |
| GROUND MOVEMENT | 0.78 | 0.76 | 0.77 | 1442 |
| INCORRECT OPERATION | 0.57 | 0.44 | 0.50 | 9 |
| LOOSE BOLTS | 0.75 | 0.35 | 0.48 | 17 |
| LOOSE FITTING | 0.00 | 0.00 | 0.00 | 12 |
| MATERIAL DEFECT | 1.00 | 1.00 | 1.00 | 2530 |
| MATERIAL OR WELDS/FUSIONS | 1.00 | 0.99 | 0.99 | 2686 |
| NATURAL FORCES | 0.71 | 1.00 | 0.83 | 10 |
| NOT PROVIDED BY REPAIR CREW | 1.00 | 0.97 | 0.98 | 30 |
| NUT FOLLOWER | 0.88 | 0.79 | 0.83 | 28 |
| NUTS LOOSENED | 0.69 | 0.59 | 0.63 | 909 |
| OTHER | 0.50 | 0.20 | 0.29 | 94 |
| OTHER OUTSIDE FORCE DAMAGE | 0.50 | 1.00 | 0.67 | 1 |
| PREVIOUS PIPE WALL DAMAGE | 1.00 | 0.92 | 0.96 | 13 |
| PULLED OUT | 0.00 | 0.00 | 0.00 | 3 |
| RUBBERS | 0.62 | 0.41 | 0.49 | 39 |
| SEAL LEAK | 0.92 | 1.00 | 0.96 | 12 |
| SHRUNKEN GASKET | 0.85 | 0.79 | 0.81 | 14 |
| STILL IN SERVICE | 0.00 | 0.00 | 0.00 | 8 |
| THIRD PARTY DAMAGE | 0.45 | 0.29 | 0.36 | 51 |
| THREAD LEAK | 0.00 | 0.00 | 0.00 | 2 |
| TIGHTENED | 0.00 | 0.00 | 0.00 | 9 |
| TREE ROOTS | 0.00 | 0.00 | 0.00 | 2 |
| micro avg | 0.92 | 0.92 | 0.92 | 17123 |
| macro avg | 0.55 | 0.53 | 0.53 | 17123 |
| weighted avg | 0.91 | 0.92 | 0.91 | 17123 |

**Confusion Matrix:**




Confusion matrix shows, counts outside of the two diagonal lines as mixed classification labels.

**Feature Importance:**

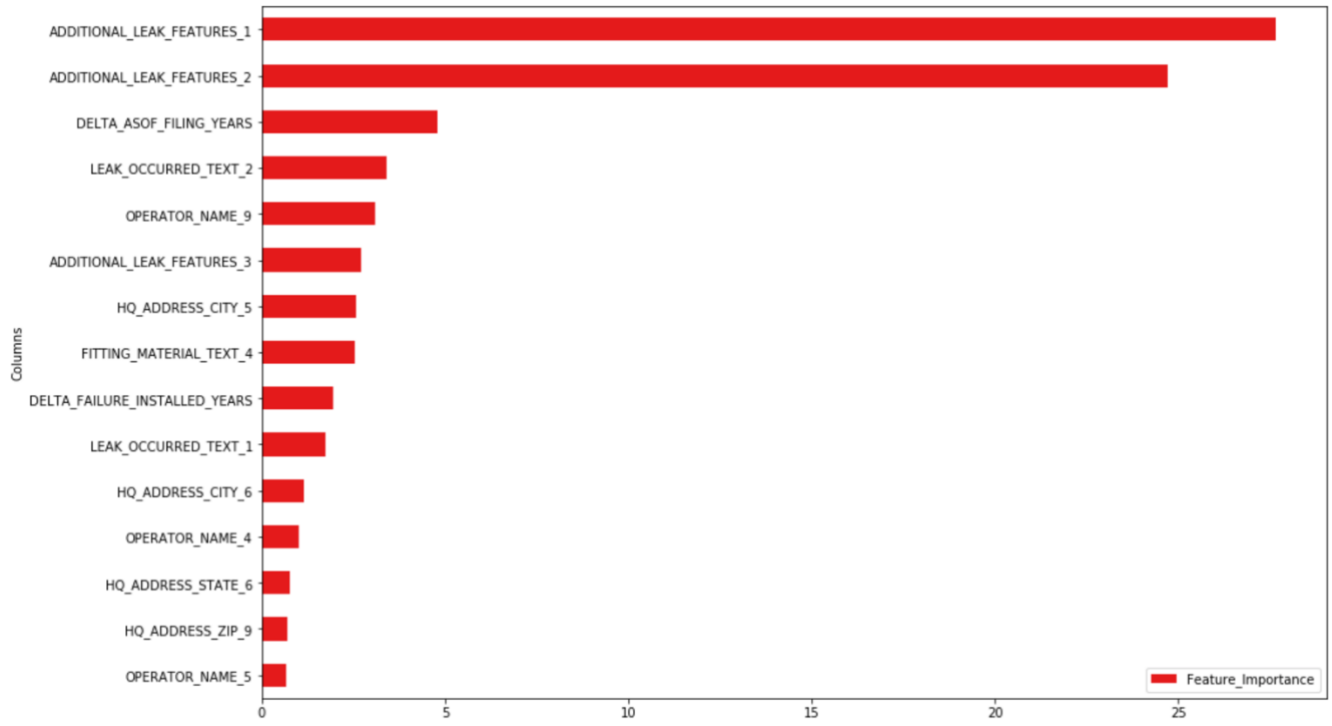
Classifier shows top 15 important features which contributed 79.9% in predicting leak cause in the pipeline with 91.6 accuracy.

```
Top 15 Features for Classifying Leaks in Pipeline:
 Feature_Importance    79.525348
dtype: float64
```

| Columns | Feature_Importance |
|---|---|
| ADDITIONAL_LEAK_FEATURES_1 | 27.665679 |
| ADDITIONAL_LEAK_FEATURES_2 | 24.703340 |
| DELTA_ASOF_FILING_YEARS | 4.783924 |
| LEAK_OCCURRED_TEXT_2 | 3.396982 |
| OPERATOR_NAME_9 | 3.096549 |
| ADDITIONAL_LEAK_FEATURES_3 | 2.719892 |
| HQ_ADDRESS_CITY_5 | 2.588522 |
| FITTING_MATERIAL_TEXT_4 | 2.554088 |
| DELTA_FAILURE_INSTALLED_YEARS | 1.934990 |
| LEAK_OCCURRED_TEXT_1 | 1.751176 |
| HQ_ADDRESS_CITY_6 | 1.153430 |
| OPERATOR_NAME_4 | 1.029005 |
| HQ_ADDRESS_STATE_6 | 0.767364 |
| HQ_ADDRESS_ZIP_9 | 0.716673 |
| OPERATOR_NAME_5 | 0.663734 |

Distribution of top 15 features used in classifying pipe leaks:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8403617240>
```




**Model A (one-hot):**

Using same hyper-parameters, Model A resulted in comparable model accuracy 90.4% and leak cause classification prediction. However, it took longer run time and required way more more features compared to Model B.

**Conclusion:**

- Random Forest using Binary Encoded Data (Model B) classified top 15 causes for leak with 91.6% accuracy.
- Random Forest using one-hot encoded data (Model A) classified top 15 causes for leak with 90% accuracy.
- Model B run time was faster than Model A due to significant difference in column sizes (136 Vs 3,912)
- Model B used top 15 features to predict 79.9% of the leak classification while Model A classified only 48.8%.
- Random Forest using Binary Encoder (Model B) was preferred.
- Important Features which classified leak cause correctly were:
  - Excavation damages
  - Natural forces
  - Welding defects
  - Leak Through the Body or Seal
  - Time elapsed between 'Installation' and 'Failure Date'.

- Time elapsed between 'As of' and 'Filing Date'
- Pipeline Operator

**Future Improvements:**

**Existing Model**: Even though Model B accuracy score is 91.6%, there were areas of mixed classification which could be improved in future models.

> Example: Elastomer Failure had 7,642 examples where Leak was identified due to elastomer failure. However, it also had 144 examples of construction/installation defects, 191 examples of ground movement defects and 162 examples of Nuts Loosened defects under the same class.
>
> We would need to check with SME's to see if elastomer failure could be caused due to these examples.
>
> Moreover, exploratory analysis showed significant reasons for leak caused due to equipment failure. We would need equipment data to understand which equipment features led to fitting failure in the past, which is currently not known.

**Apply Other Advanced Models**: I would like to run deep learning model on the same dataset and compare leak classification result. Also, use predictive modeling to estimate when a mechanical fitting could fail based on the number of years of operations.

**Build Dashboard**: I would like to build a real-time dashboard on key performance indicators such as expected failure date, operator and manufacturer details, and likely reason for leak cause, to help clients avoid any operational risks, early-on.

**---End of Project Report---**